Arndt Britschgi

Anything Goes,
No Paradox Follows

Introductiones

Contributions to Philosophical Analysis

Arndt Britschgi

Anything Goes, No Paradox Follows

A Free-Will Investigation into Newcomb's Paradox

Philosophia

Die Deutsche Bibliothek
CIP-Einheitsaufnahme
Der Titelsatz für diese Publikation
ist bei der Deutschen Bibliothek erhältlich

Die vorliegende Arbeit
wurde von der Philosophischen Fakultät
der Universität Zürich im Sommersemester 2005
auf Antrag von Prof. Dr. Rafael Ferber
als Dissertation angenommen.

ISBN 3-88405-701-8

Printed in Germany 2006

I thank Prof. Dr. Rafael Ferber for his support
and encouragement
throughout the process of completing this work.
Thanks also to Dr. Georg Brun for his intelligent advice
on many importanrt details;
and to the Paul Schmitt Gedächtnisstiftung for their
financial assistance.
For whatever it may be worth, I dedicate the book to Monika.

Contents

Introduction 9
1. General 9
Newcomb's Paradox (Standard Version) 9
Short History, Comments 10
2. "A Beautiful Problem" 15

Chapter One: How Should We Choose, and Why? 19
1. Exposition 19
2. Two Principles of Choice: MEU and DP 25
3. Mackie: The Direction of Causation 37
4. Backwards Causality: The Semantic Objection 44
5. Nozick: The Illusion of Influence 46
6. Comments 50
Bar-Hillel and Margalit (1972) 50
Martin Gardner (1973/-74/-86) 52
George Schlesinger (1974) 52
James Cargile (1975) 53
Richard Grandy (1977) 54
Michael Stack (1977) 54
Terence Horgan (1981) 55
S. L. Hurley (1994)/Nozick (1993) 57
Two Reflections 58
One: Abracadabra 58
Two: Why necessarily opaque? 59
7. End Remark to Chapter One 59

Chapter Two: Predictor P, Absolute Infallibility 61
1. Three Possibilities? 61
P Is Infallible (1) 61
A free choice 62
Naive (Absolute) determinism 74
A Nearly Absolute Infallibility (2-3) 80

Chapter Three: Statistical/Decision Theoretical Addendum 89
1. Attempted Positioning 89
2. Expected Utility 99
Is the Decision in NP Unstable? 113
Comment 118

Chapter Four: The Agent W, a Free Decision 119
1. Will Versus Matter 119
2. Generating a Decision (Modal Approach) 123
3. Possibilities (Model Theoretical Approach) 152
Possible Worlds: Two Concepts 153
Ways the World Might Have Been 157
Modality – Model Theory 160
4. End Remark to Modalities 166

Chapter Five: Backward Causality (Logical Aspect) 169

Chapter Six: Summary 173

References 175

Introduction

1. General

Newcomb's Paradox (Standard Version)

A treatise on Newcomb's Paradox hardly stands out in terms of the originality of its subject. Since Robert Nozick introduced it in an essay in 1969 numberless authors have made numberless efforts to produce a solution. If I choose to analyze it yet again it is because of what the analysis reveals implicitly about a world of ever unresolved contradictions; my own attempt at a solution (if I make such an attempt – I tend to think that there is neither a solution nor a problem) will, by far, be more fruitless than the ones that I examine.

As standard version I will use one taken from A. Gibbard and W.L. Harper (1978: 180-81). An agent [W] faces two boxes, one transparent [A] and one opaque [B]; the transparent box contains a thousand dollars [$1000]. The agent can perform the action of taking the contents of the opaque box only [Only-B], or the contents of both boxes [AandB], and knows that a predictor [P] has already placed a million dollars [$M] in B if he predicted Only-B and nothing if he predicted AandB. The agent also knows that the probability of a correct prediction in either case is close to one.[1] Assuming that the agent wants to attain the largest possible amount of money, which of the actions is the better choice?

I have written on the problem in two earlier papers, and will go on using some of the results in the study that follows. In particular, I will not take the concept of a quantum-mechanically driven leap, and, in direct consequence, an immanent cause within the causal chain, into further account, as I consider Ted Honderich's (1993) case against this view valid. Subjecting an action or a choice to what in the end is coincidence will certainly not make them free. I do not deny, in this context, the possibility of an immanent cause as such, and quite specifically a quantum-mechanically conditioned immanent cause; yet I think it should be seen more in the sense sir John

[1] The denotations in square brackets are my own. As Gibbard and Harper indicate a male predictor, I will maintain the masculine gender for my P throughout the text – to be consistent, likewise for my agent W. I will frequently be using the short form NP for Newcomb's Paradox/Newcomb's Problem.

Eccles (1994) suggests, i.e. we should move beyond the purely neural to a microneural level, where it is not so much (or at the most in a derived sense) a question of a causal leap, as of the effecting of a decision and its corresponding action in itself.

I will also consider the issue of to what degree the predictor P, on a neural level, could observe the course of events as settled: I will assume that P, theoretically speaking, could overlook these events and their corresponding mental states and, based on the knowledge thus gathered, make a prediction of the agent's future choice (which in itself does not mean that he could predict it with certainty). This is due to the thought-experimental character of the Newcomb problem, which I explored in the previous papers and also return to briefly below.

Throughout the text I will make use of the sources rather freely; besides what they say I will work on what I think their concepts suggest or necessarily must lead to, without always pointing out a clear difference. In cases where the ideas are clearly not expressed in the original texts, explicitly or implicitly, I will put my observations in square brackets.

Short History, Comments

The problem first appears in Nozick (1969), as I have mentioned. It was widely popularized by Martin Gardner in his column in the *Scientific American*. In his essay Nozick claims that it originally stems from the physicist William Newcomb (from whom it has its name: Newcomb's Problem or Newcomb's Paradox); Nozick himself had heard about it from a friend in 1963. He labels it a "beautiful" problem, which, somewhat melancholically, he regrets is not of his invention.

Nozick's (1969) version differs from our standard version above in that it adds a further specification (in the form of a footnote, or, more precisely, an appendix): If the predictor foresees a "random choice" (i.e. a choice determined by some randomizing procedure in the vein of flipping coins, or similar), then he does not place the money in the second box regardless of which of the actions W performs. In other words, Nozick considers any random choice automatically equal to the two-box choice AandB. He does not give an explanation as to why such a condition be required or acceptable; the

way he represents the matter it is also not quite clear whether the addition originates from him or was included in the version as he heard it.

Nozick's version remains unclear on two further, central points. "You know that this being [P] has *often correctly* predicted your choices in the past, and . . . that this being has *often correctly* predicted the choices of other people [in this particular situation]." (Nozick 1969: 115, emphases added) Does this mean that P has often made an accurate prediction, although on some occasion or some occasions he has failed, or has he often made a prediction and every time an accurate one? The second unclear point concerns the general time frame. We know the sequence P's Prediction-Distribution-W's Choice, we do not know where in this sequence W is presented with the problem. At what exact point does the agent get to know his coming assignment?

Maya Bar-Hillel and Avishai Margalit (1972) defend the one-box choice Only-B. In their version the ambiguity of "often correctly predicted" is resolved, as they unmistakably state that P so far has never made a wrong prediction (295). Also regarding the second point they seem to offer a clarification: "The Being has just now (or an hour ago, or a year ago...) made his prediction and his move." (ibid.: 295) This at least implies a time sequence Prediction-Distribution-*Initiation*-Choice.

Other notable comments are J.L. Mackie (1985b) and the essay by Gibbard and Harper (1978) referred to above. Richmond Campbell (1985) has written a helpful overview, whereas R.M. Sainsbury (1988) makes an extensive, not very successful attempt to dissolve the paradox. As far as I know, no one as yet has proposed a generally recognized solution. As Nozick (1969) affirms, the distribution between advocates for one or the other choice is surprisingly even, among philosophers as well as non-philosophers; typically, the opinions on both sides tend to be inflexible. The feedback on a website inquiry conducted in the late 1990's (Lin 1997) shows a 61 percent preference for Only-B among three hundred participants. Over time, the one-box solution seems to have slightly gained an edge.

Max Black (1983) maintains Nozick's condition regarding a random choice in what he considers the (or a) basic version of the problem.

Black presupposes that P, in the case of the present agent W, has so far always made an accurate prediction, whereas in the case of other agents "similar" to the present agent he has almost always made an accurate prediction; hence one can "confidently" expect a correct prediction now as well.

To make the situation more realistic, Black goes on to develop NP into the so called "Museum Game," which includes some distinctive modifications in relation to his own basic version. The modifications, shortly summarized, are as follows. First, P no longer makes the claim of absolute infallibility; as of now his prophecy is only almost certainly correct.[2] Secondly, the predictability of the choice is to be made plausible to the agent by convenient changes in the sums involved, by introducing entrance fees and questionnaires about the agent's character, etc. Thirdly, the random choice condition is canceled, or at least no longer mentioned to the agent. Fourthly, the Museum Game should be created in such a way that it provides a reasonable explanation of why it is being played at all, and, fifthly, it should provide the agent with some grounds to do his share quite earnestly. Except the first and third adjustments, this seems merely trivial (cf. footnote 2). The changes strictly aim at stressing that the game is being fair, so that the agent for his part will make an honest choice. What the Museum Game shows, then, is that Black has missed a point: Newcomb's Paradox is notably a thought experiment and not a game you could imagine being played realistically. The situation it describes would not be credible at all under any kind of lifelike circumstances. Among the traits of a thought experiment is fundamentally that all the details and the rules are set and mutually accepted – they are valid just because we all agree that they are valid, and simply picture them as being valid. (Only logical contradictions are excluded in these cases.)

Mackie (1985b) interprets P's reliability somewhat sharper as he states that P so far has *always* made a true prediction; in his account a random choice is not considered. Mackie also recommends modifications: He would rather have the $M in B reduced to $10,000, basically because the $1000 in A thus grows to be a more significant additional amount. In doing so Mackie commits the same mistake as

[2] I cannot see in what sense this is a modification to Black's basic version.

Black above. In its status of a thought experiment NP establishes as a fact that the agent wishes for the largest possible amount. That the reward objectively seen be more or less appealing does make a technical difference in driven decision theoretical analyses (specifically in what we know as the "regret theories" – we will also see below, in connection with such theories, that Nozick could have had his motives when he fixed these sums precisely as he did); however, given the fact that success/failure in this context is reduced exclusively to more or less money, the actual difference in sums does not in practice, in the way Mackie proposes here, have a real influence. One dollar more or less already makes the whole difference between a total victory and absolute defeat.

Like other authors, Mackie too takes into account the possibility of repeated trials. I find this badly in accordance with the experiment as described – a point that Mackie recognizes as he sums up his results. The way I see it, what is finally at stake is how to choose in this particular and given situation, just this once, so as to gain as large a sum as possible; except for what they might reveal about P's wondrous predictive powers (which in any case should be infallible, or very close), earlier trials should have no impact on the choice at hand right now. There is no room to think, for instance, that by repeating the same choice a number of times you could deceive P into thinking that in fact it must be linked to some inherent trait of your character, something you later would make use of by a sudden change of course. The same goes equally for every kind of trickery or feints, like sleight of hand, mentally faked, not-actually-carried-out decisions, etc.; and equally also for statistical, game-theoretical levelling-out effects through an indefinitely increased number of repetitions. If you choose Only-B, then P has foreseen your intention and also acted in correspondence with his knowledge of your choice; if on the other hand you choose AandB, then P has foreseen that intention and once again acted in correspondence with his knowledge. These are the sole two possibilities, you are convinced of this the moment when you choose. (Unless you go for a notorious "random choice" in spite of all, knowing that P has foreseen this and left the second box empty: that gives a further possibility where what you score is nothing.)

In Gibbard and Harper (1978), our model version, the prediction is described as "highly reliable;" as we have seen, they give a probability of close to one for a right forecast. Sainsbury (1998) states that P thus far has made absolutely no mistake. Neither he nor Gibbard and Harper maintain Nozick's random-choice condition. We should notice once again what Nozick (1969) says about "the being" (Nozick logically is the starting point for every subsequent version): in the past he has often correctly predicted the choice of the present agent and that of others in an identical situation, and, as far as the present agent knows, never made a mistake. The prediction will almost certainly be correct on this occasion as well. So apparently, in the end, Nozick too assumes "often" to mean, not that P has often succeeded in his prediction, although it sometimes did not turn out right, but that there has been numerous trials without there ever occurring a mistake on the part of P. We see that none of the versions states a logically construed, absolute infallibility, whereas most of them support a 100 percent accuracy in practice. As to the general chronology, all authors seem to agree: P makes his prediction, distributes the money, and only then W makes his decision. Regarding at what point in this chain W is introduced to the problem all versions remain more or less vague. I will assume that the introduction happens after the prediction and the distribution, since this is the only way in which the problem presents a reasonably consistent structure (cf. further below).[3]

I offer the opinion that the idea of Newcomb's Problem consists in the fact that we are actually *certain*, i.e. in this sense *know beforehand* what amount of money we will get through one or the other choice, quite regardless of whether we work with probabilities, logical infallibility or whatever conceivable concept there might be; regarding this crucial aspect Bar-Hillel's and Margalit's presentation is perhaps the most illuminating: "...you are, *for some reason, almost sure* that if you will take both boxes you will end up with $1000, whereas if you will take just the covered box, you will end up with a million dollars." (1972: 295, emphasis added) As we choose we may

[3] Regarding the random-choice condition, I will consider it as I advance a little in the spirit Nozick does himself: as an additional detail which should not be wholly ignored. As of a certain point it gains decisive weight, however.

be completely convinced that AandB will produce $1000 (or simply *a smaller amount*), Only-B a million (or simply *a larger amount*); on the other hand, we also know that, one, the $1000 are already in A (we see them) and, two, the million already in B or, alternatively, already removed from B (P has already predicted our choice and acted). On these conditions, which of the two actions Only-B and AandB is more rational? That is the essence of the problem. Newcomb's Paradox, as we have seen, is a thought experiment, it has to be treated in this capacity throughout.

2. "A Beautiful Problem"

Why does Nozick find the problem so particularly beautiful? It is an astonishing characteristic of this "mock puzzle," Black (1983) claims, that the extensive and subtle discussion it inspired as yet has led to no generally acknowledged solution;[4] Black further confirms that Nozick's original essay triggered a flood of letters with suggestions for solutions and comments. At first sight Newcomb's Problem no doubt looks like a paradox in an authentically traditional sense: under the given circumstances there are two, and only two possibilities, both of which lead to an obvious contradiction; at the same time it seems just as obvious to everybody involved that either one or the other choice is the solely better solution – the only problem is to convince the rival side of this plain fact. We lack an ultimate, once and for all persuasive argument. Some categorically defend the one-box choice, others just as categorically AandB; this confrontation is the cause of a debate that never leads to a clear narrowing of the positions, but which on the other hand summons a range of basic philosophical questions and lets them stand out in a heightened light. In this aspect Newcomb's Paradox is a prolific problem: it fuels analyses on time and space, determinism, free will and the conception of a choice in general. Compatibilism and epiphenomenalism become points of strong contention. Formal-logical, moral- and le-

[4] As no reliable dates to the contrary have appeared since then, we can go on accepting Black's claim as corresponding to the actual state of affairs.

gal-philosophical, theological notions and theories are thrown into the debate and tested.

Moreover, there are the poetical qualities of the evoked picture. Here the eminent predecessor and paradigm is Zeno: like his famous paradox about Achilles and the tortoise, which in itself makes up a beautiful fable, Newcomb's Problem also offers rich opportunity for creative flare. Without compromising any of the premises you can expand the situation in a great number of ways, and, what is more, vivid images of W facing the ominous pair of boxes, the prophet P an enigmatic, superhuman, transcendental presence ever watchfully in the background, literally force themselves upon you. Up to this point NP displays thoroughly classical virtues; on a further point, however, it falls short of its great model.

Whereas Zeno's paradoxes come out theoretically "clean," in the sense that all the facts are present in the primary situation, Newcomb's Problem does require what I would call a made-up precondition: we "know" something that does not follow from the premises alone, or in some other way would be self-evident. For it is evident (whatever reason there may be to make it so) that Achilles will catch up with the turtle, or that a runner, given time, will reach the goal line of the track (let us say that, in the view of crushing empirical evidence, we cannot doubt these facts); and it is just as evident, or those of us who find it evident *experience* it as just as evident, that the runner first must reach the half point mark of the track, then the half point of the half point, and so on *ad infinitum*, so that he actually never makes the goal line after all, and we get the paradox (similarly in the case of fast Achilles and the turtle). So the runner and Achilles simply need to get on their way, and there we have the paradox; there is no call for anything added. Why there should be a being P who knows exactly how we will choose on the other hand does not seem evident at all: we even tend to doubt that very much. To "know" as much we must assume something that does not strictly follow from the original situation as we had it – a not uncontroversial theory (determinism), a not very plausible statistical information (P's 100-percent record up to date), or the like – although it is clear that *if* there exists such a being, and *if* that being distributes the money in the way that we are told, *then* we have the paradox that AandB is more rational than Only-B, and at the same time Only-B

more rational than AandB. In any case, we cannot simply go ahead and choose and that already would create a paradox; we get the paradox by supposing some implied, artificial extra.

More precisely we could say that Zeno's paradoxes have a consequential form: because on one hand X and on the other hand Y, and since X and Y are mutually exclusive, we have a paradox. Analogously, Newcomb's Problem has a conditional form: if we know that P always makes a correct prediction, and if, under this condition, we have to choose between Only-B and AandB with the intention of getting as much money as possible, then both options are more rational than the other and we have a paradox – if the condition, nonetheless, were *not* the case, we would not have one.[5] But there is no case, no rationally conceivable, counterfactual situation whatsoever in which 'Achilles and the Tortoise' or 'The Runner' are not paradoxical (as long as we are not prepared to change the very basis of our conception of reality, in which case we find ourselves before a wholly different, ontological question), the paradox here is inherent in the situation as it is. To use the imagery of NP, let us conclude as follows. Had Nozick invented Newcomb's Paradox himself, as he wishes in the footnote referred to above, then we could maybe have given him $1000 in prize money; if on the other hand he had been the first to think out one of Zeno's paradoxes, then he clearly would have earned a million – the beauty here is on a wholly different level. Or, as long as no convincing resolution is in sight, for inventing one of Zeno's paradoxes we could have afforded him the all-time jackpot: one million dollars and the $1000 as a bonus.

My critique will center on the weak point indicated above. In the first chapter I will expand on the basic features of the problem, present two principles introduced by Nozick to justify Only-B and AandB, respectively, and analyze some further suggestions for a solution, proffering the thesis that the unsolvability of Newcomb's Problem derives from a contradiction in its structure. In Chapter

[5] And here it is not a matter of whether we believe the premise of P's infallibility or not, but precisely that this premise stands in a particular relation to the rest of the preconditions.

Two I try to show how a statistically immaterial error rate in P's prediction inevitably equates to an absolute infallibility and what the consequences of an absolute infallibility amount to; Chapter Three is a statistical, decision-theoretical complement to the previous two, and Chapter Four in its first part an analysis of the corresponding effects on the agent W and his particular choice. The second part of Chapter Four is a modal logical inquiry into the possibility of a choice in general. In the fifth chapter I will consider a logical aspect of the principle of backwards causality, which may be of weight to the investigation as a whole. Throughout the investigation further questions of philosophical interest will come up; insofar as they refer to the key subject I will discuss them in the respective sections of the text. I will sum up the results in the sixth chapter. I maintain that Newcomb's Problem seems a paradox only because its premises are inconsistent between themselves. It simultaneously assumes an infallible P (who, as such, is not contradictory) *as well as* a free choice (which *in itself*, I believe, is not contradictory), and both of them *together* constitute a contradiction – the fact that contradictory premises lead to a contradiction in the conclusion is not paradoxical, but, by virtue of the *ex falso quodlibet*, only logical.